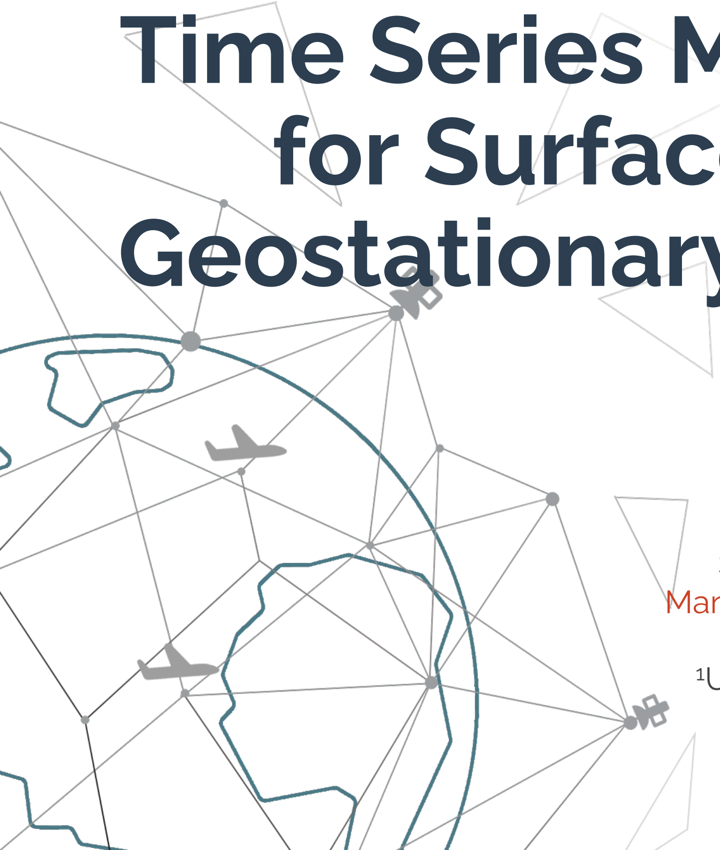


Time Series Machine Learning Methods for Surface PM2.5 Estimations Using Geostationary Satellites and Numerical Weather Models

Wednesday, 26 January 2022; 3:45 - 5:00 PM CST

George Priftis[1], **Aaron Kaulfus[2].** Muthukumaran Ramasubramanian[1], Shubhankar Gahlot[1], Iksha Gurung[1], Manisha Khatri[1], Peiyang Chaeng[1], Manil Maskey[2], Rahul Ramachandran[2], Sundar Christopher[1], Haeyong Chung[1]

[1]University of Alabama in Huntsville, [2]NASA Marshall Space Flight Center

102nd American Meteorological Society Annual Meeting,
21st Conference on Artificial Intelligence for Environmental Science

Introduction

Tropospheric aerosols impact weather and climate through scattering and absorption of radiation while also degrading air quality

- $PM_{2.5}$ is inhalable and impact respiratory and cardiovascular functions (Shaddick et al., 2020).
- Direct measurements of surface $PM_{2.5}$ concentrations represent air quality at a specific geographic location, which may not be representative of the surrounding area.

Satellite and meteorological based AOD-$PM_{2.5}$ relation facilitates contiguous $PM_{2.5}$ estimation and mitigation of spatial limitations

- Availability of large data in the last decade has enabled the use of machine learning (ML) methods to develop data-driven relations
  - Prior studies have focused on estimating $PM_{2.5}$ on polluted regions, with few studies published concentrating on the Contiguous United States (CONUS) (Di et al. 2016; Hu et al. 2017)
  - The AOD-$PM_{2.5}$ relation is highly variable spatially, therefore estimating $PM_{2.5}$ concentration on a continental scales may necessitate a large number of ML models (Chudnovsky et al. 2012).

The aim of this initiative is to:

- implement and compare different ML methods to estimate surface PM2.5 over the CONUS
- introduce a novel technique to decrease the large number of regional-based models

Data Sources

National Oceanic and Atmospheric Administration's (NOAA's) Geostationary Operational Environmental Satellite-16 (GOES-16) Level 2+ AOD

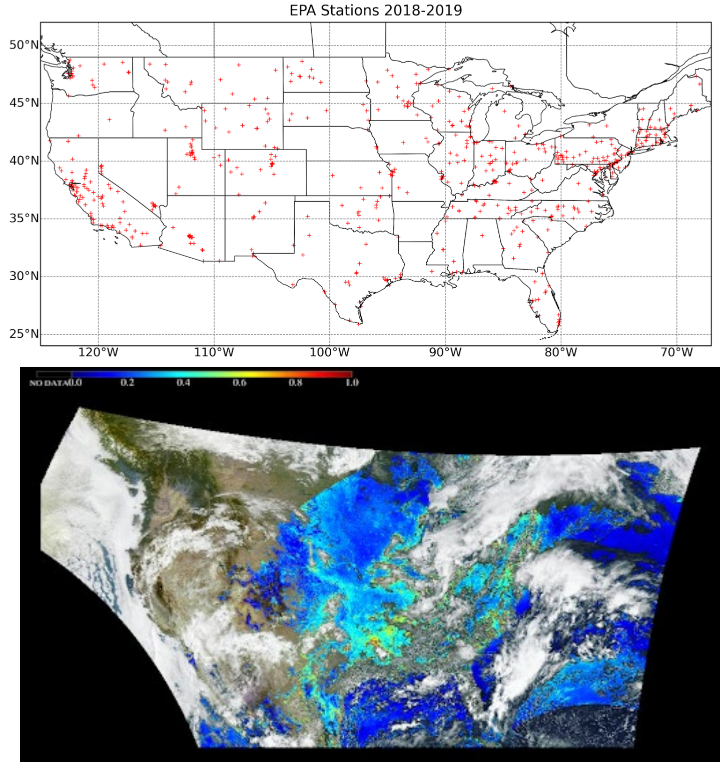
- 2 km spatial resolution
- 5 min temporal resolution

Environmental Protection Agency (EPA) PM2.5 concentrations

- 317 monitoring stations across the CONUS
- Hourly observations

NOAA's High Resolution Rapid Refresh (HRRR) model (Benjamin et al. 2016)

- 3 km spatial resolution
- Hourly analysis fields
- Parameters utilized: surface temperature (T), relative humidity (RH), u/v-wind components, and planetary boundary layer height (PBLH)



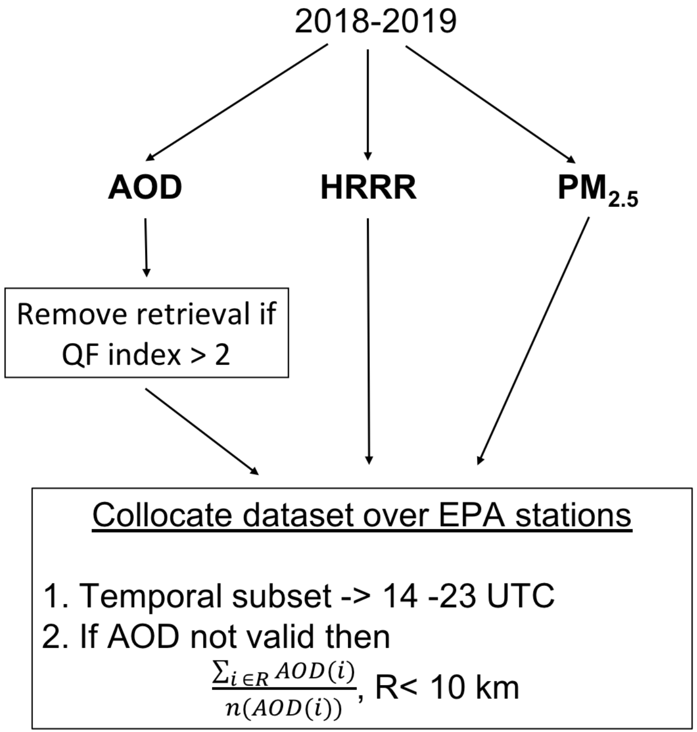
EPA stations over the CONUS with hourly PM2.5 sensors (top) and aerosol optical depth (AOD) over the CONUS for August 6th, 2019 from GOES 16.

Dataset Preparation

Machine learning dataset preparation workflow

1. Data collected for 2018-2019
2. Data for which the AOD is negative or the quality flag is larger than 2 (low quality) are removed
3. All datasets are collocation over the EPA stations at the closest time
4. Data are temporally subset to between 14-23 UTC to mitigate sun angle effects
5. When AOD is not directly collocated, pixels are extracted within a 10 km radius (euclidean distance) and averaged

The final dataset (59055), is then split into 80%/20% into training and testing sets respectively



2018-2019

AOD

HRRR

$PM_{2.5}$

Remove retrieval if QF index > 2

Collocate dataset over EPA stations

1. Temporal subset -> 14 -23 UTC
2. If AOD not valid then $\frac{\sum_{i \in R} AOD(i)}{n(AOD(i))}$, R< 10 km

Methods

To obtain a robust relation between satellite retrieved AOD and PM2.5 concentration, a multi-parametric model needs to be defined. This study takes a 2 step approach to identifying this model

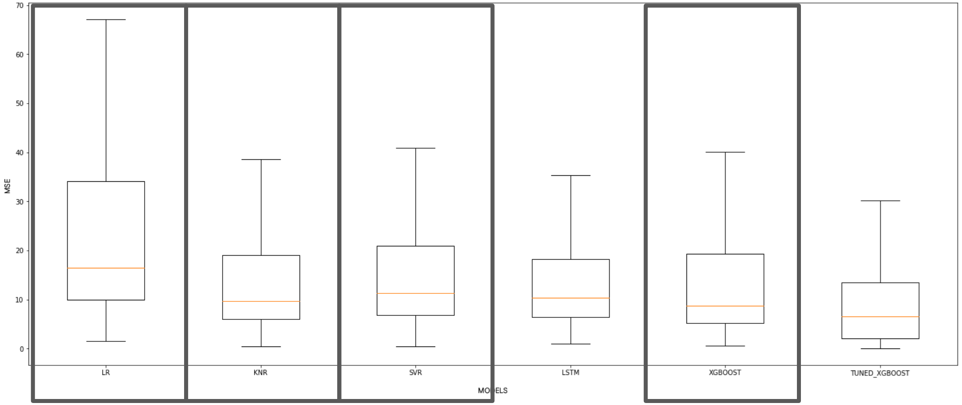
1. A comparison of different CONUS generalized ML models
    a. Linear regression (LR)
    b. k-nearest neighbour (KNR)
    c. Support vector machine (SVR)
    d. Long-short term memory (LSTM)
    e. Extreme-gradient boosting (XGBOOST).
2. Station-wise models
    a. To obtain a better representation of PM2.5 spatial characteristic station-wise LSTM models are created and trained on each individual station.
    b. Fleet training is introduced to reduce the number of models
        i. Merging individual station data into a station-wise model iteratively if the MSE after merging is decreased in comparison with the station-wise model.
        ii. Expected to lead to a decrease in the average MSE over all remaining models and provide insight into missing processes

Generalized Model Results

The Linear Regression (LR) model attains a high median MSE (18) and a large range (0-68), thus failing to capture the complex relation between the input variables and the target $PM_{2.5}$

kNR and XGBOOST have similar performance with a median MSE of approximately 10 and 8 respectively, however the latter has a smaller MSE range (0-35)

SVR and XGBOOST attain the same range of MSE (0-40), however the XGBOOST has a smaller median (8) from all the aforementioned models

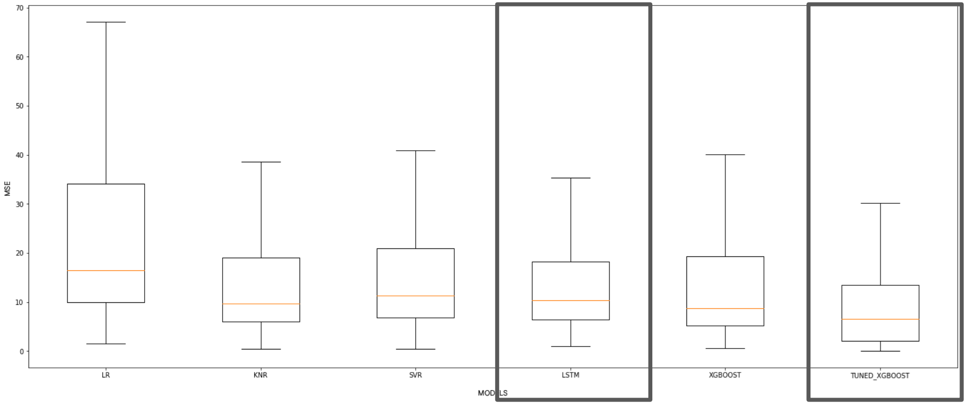


Mean Squared Error (MSE) for different machine learning models. Linear regression (LR); k-nearest neighbour (KNR); Support vector machine (SVR); Long-short term memory (LSTM); Extreme-gradient boosting (XGBOOST).

Generalized Model Results

Although LSTM is adequate to handle sequence dependence prediction problems, tree-based models such as the XGBOOST can perform better (smaller MSE) with less amount of available data.

In order to develop a robust model, the XGBOOST is tuned (e.g. number of iterations) using the learning curves as a diagnostic tool, providing the final model (TUNED_XGBOOST). This model is characterized by a median MSE of 5 and a range 0-32



Mean Squared Error (MSE) for different machine learning models. Linear regression (LR); k-nearest neighbour (KNR); Support vector machine (SVR); Long-short term memory (LSTM); Extreme-gradient boosting (XGBOOST).

Fleet Training

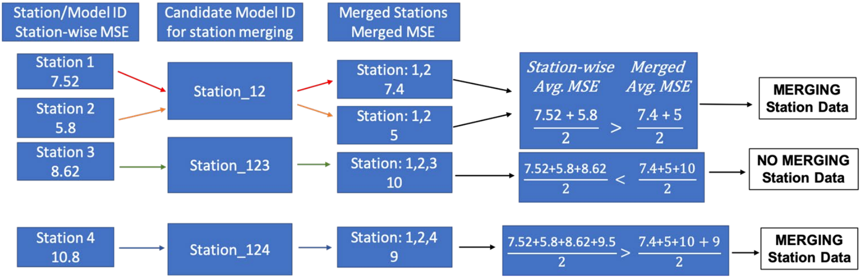
Feature selection using ensemble models has been utilized in the ML community to help with identification of best initialized for stochastic models

- The LSTM architecture was chosen as a base model to train station-wise models

A new technique, termed fleet training, is proposed to account for unknown characteristics and iteratively improve the model

- Models for each station is iteratively trained on data from other stations by merging the corresponding data.
- If the resulting model performed better (decrease in the MSE) than the original station model, then merging of the two datasets occurs.

A total of 6 iterations are performed in this case leading to a decrease in the average MSE from 9.134 to nearly 7.783 and from 317 to 129 models



Station/Model ID Station-wise MSE

Candidate Model ID for station merging

Merged Stations Merged MSE

Station 1 7.52

Station 2 5.8

Station_12

Station: 1,2 7.4

Station: 1,2 5

Station-wise Avg. MSE *Merged Avg. MSE*

$\frac{7.52 + 5.8}{2} > \frac{7.4 + 5}{2}$

MERGING Station Data

Station 3 8.62

Station_123

Station: 1,2,3 10

$\frac{7.52+5.8+8.62}{2} < \frac{7.4+5+10}{2}$

NO MERGING Station Data

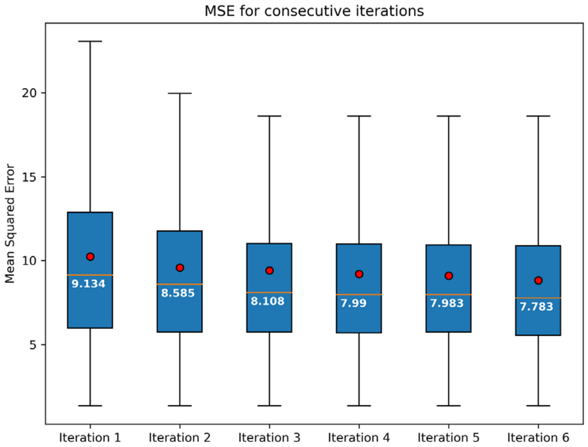
Station 4 10.8

Station_124

Station: 1,2,4 9

$\frac{7.52+5.8+8.62+9.5}{2} > \frac{7.4+5+10 + 9}{2}$

MERGING Station Data



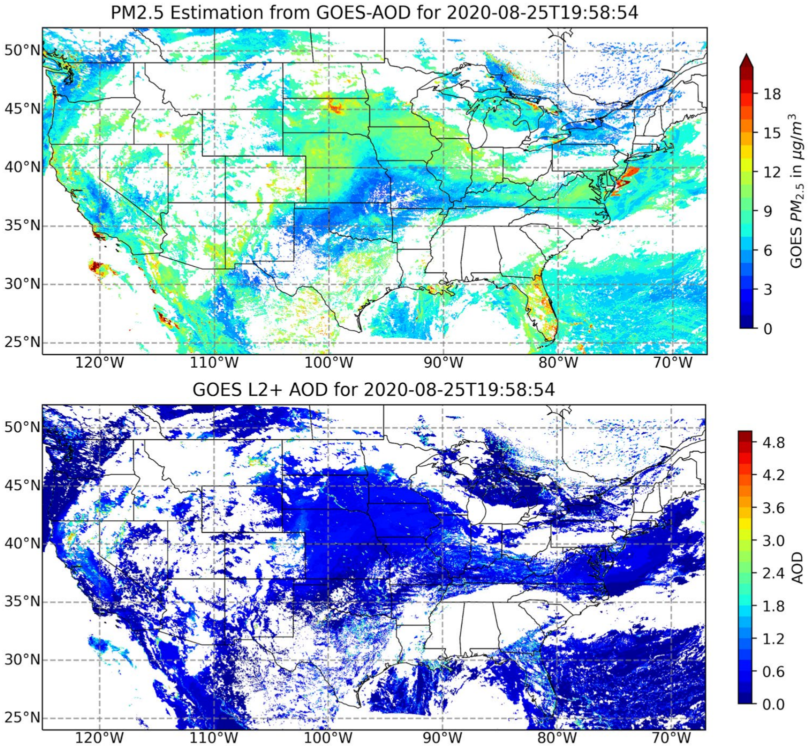
Example of model architecture for the fleet model (top) and MSE for resulting iterative modeling training (left)

PM2.5 Estimation Use Case

The generalized model was used to calculate average $PM_{2.5}$ concentrations for each grid cell on August 25, 2020 at 19:58UTC during the active California wildfire season.

- Maximum estimated $PM_{2.5}$ is 44 μm m^{-3} and the 85th percentile has a value of 9.06 μm m^{3}.
- Several hotspots are prevalent over the CONUS:
  - Florida
  - South Dakota
  - Northern California

Corresponding AOD retrievals from GOES-16 are also elevated over Texas and Northern California. However, low AOD is observed over South Dakota



PM2.5 Estimation from GOES-AOD for 2020-08-25T19:58:54

GOES L2+ AOD for 2020-08-25T19:58:54

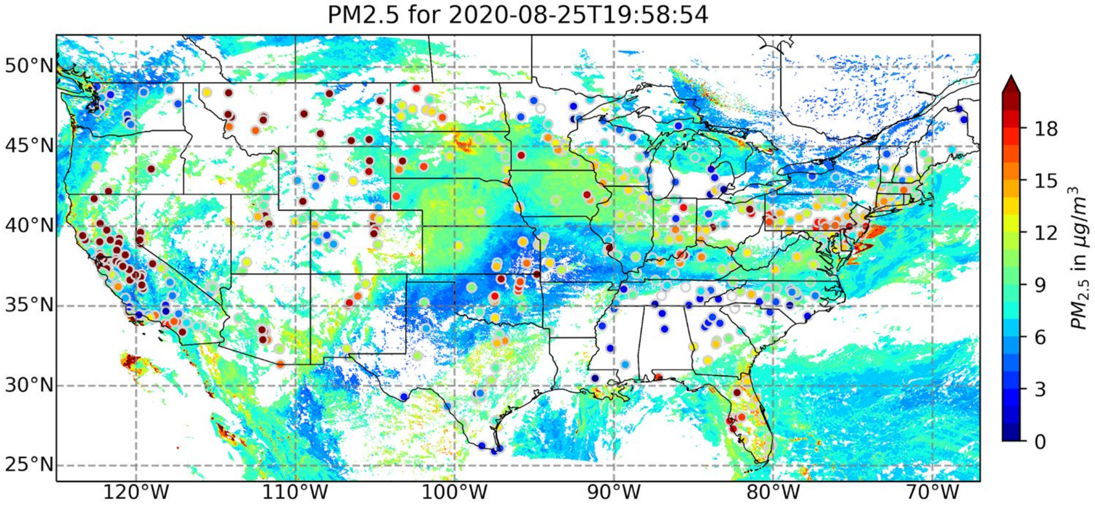
$PM_{2.5}$ estimation from the generalized XGBOOST model over CONUS for 2020-08-25 at 19:58:54 UTC. (top) and Aerosol optical depth retrieval from GOES-16 for 2020-08-25 at 19:58:54 UTC (bottom).

PM2.5 Estimation Use Case

$PM_{2.5}$ measurements from the EPA ground-based stations have a maximum of 112 µm m^{-3} and 85th percentile of 20 µm m^{-3}

Elevated concentration of $PM_{2.5}$ over the California fires is not well captured by the model estimates which might be attributed to:

- low number of samples of extreme $PM_{2.5}$ in comparison to the total number of training samples
- to high AOD retrievals aloft for very thick smoke



$PM_{2.5}$ estimation from the generalized XGBOOST model over CONUS for 2020-08-25 at 19:58:54 UTC. Scattered points correspond to $PM_{2.5}$ concentration obtained from the EPA ground-based stations at 20:00 UTC. $PM_{2.5}$ values are capped at 20 µm m^{-3}.

Summary

Nonlinear relationships between emissions, meteorology, and chemical pathways makes estimating PM2.5 concentrations difficult with highly specialized numerical methods widely adopted

Study seeks to apply novel, data driven, machine learning methods for the estimation of hourly PM2.5 concentrations

Fleet training using LSTM models is introduced as a means to identify key relationships modulating PM2.5 concentrations

- This methods demonstrates that for continental scale estimations, the number of site specific models can be significantly reduced while reducing error

XGBOOST model are the best performing while also having the advantage of requiring a lesser amount of data and processing required

- A generalized XGBOOST model predicts surface PM2.5 well a continental scales except at very high AOD's and when aerosols are aloft

Thank you.